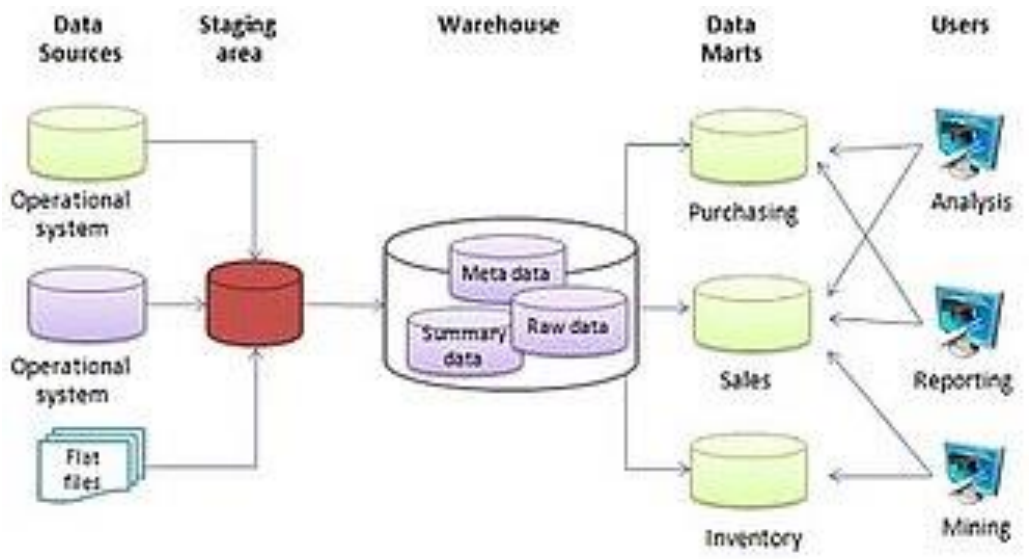
Data Warehousing: Basic Concepts

Hannah Andrews and Angela Hughes

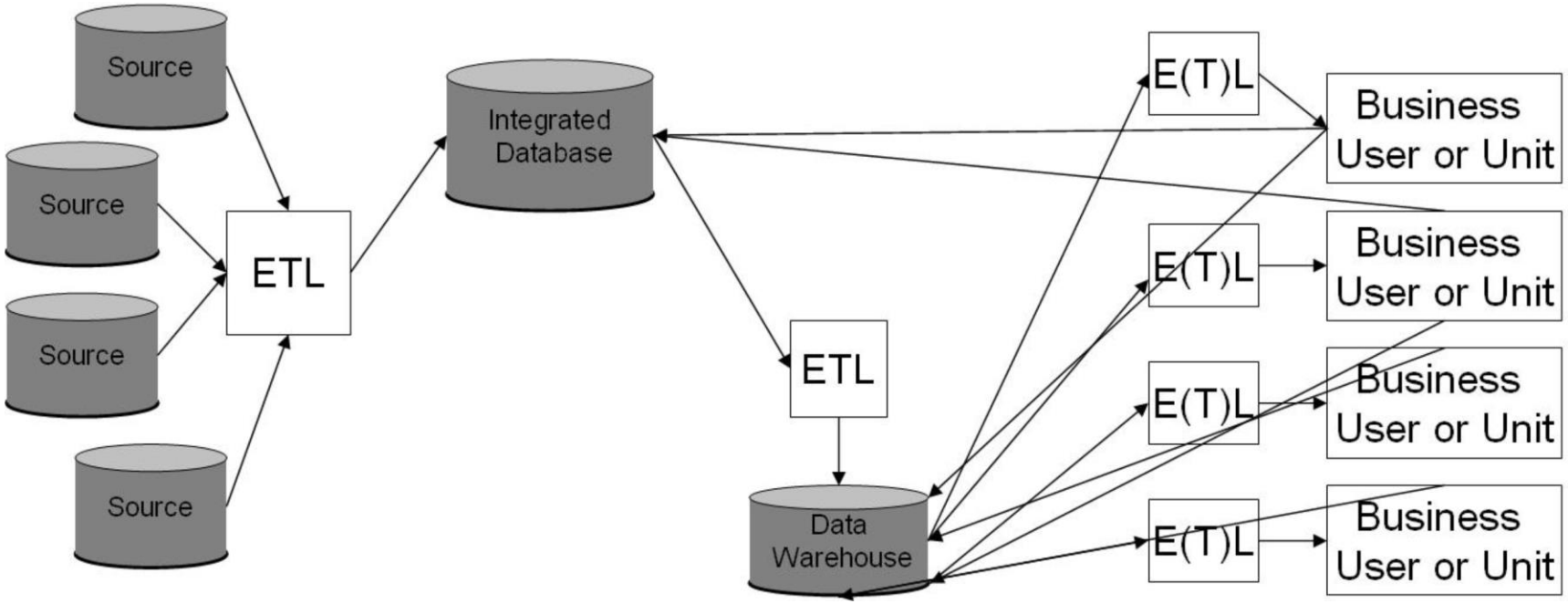
"A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decision-making process."—W. H. Inmon

- Subject-Oriented
- Integrated
- Time-variant
- Nonvolatile



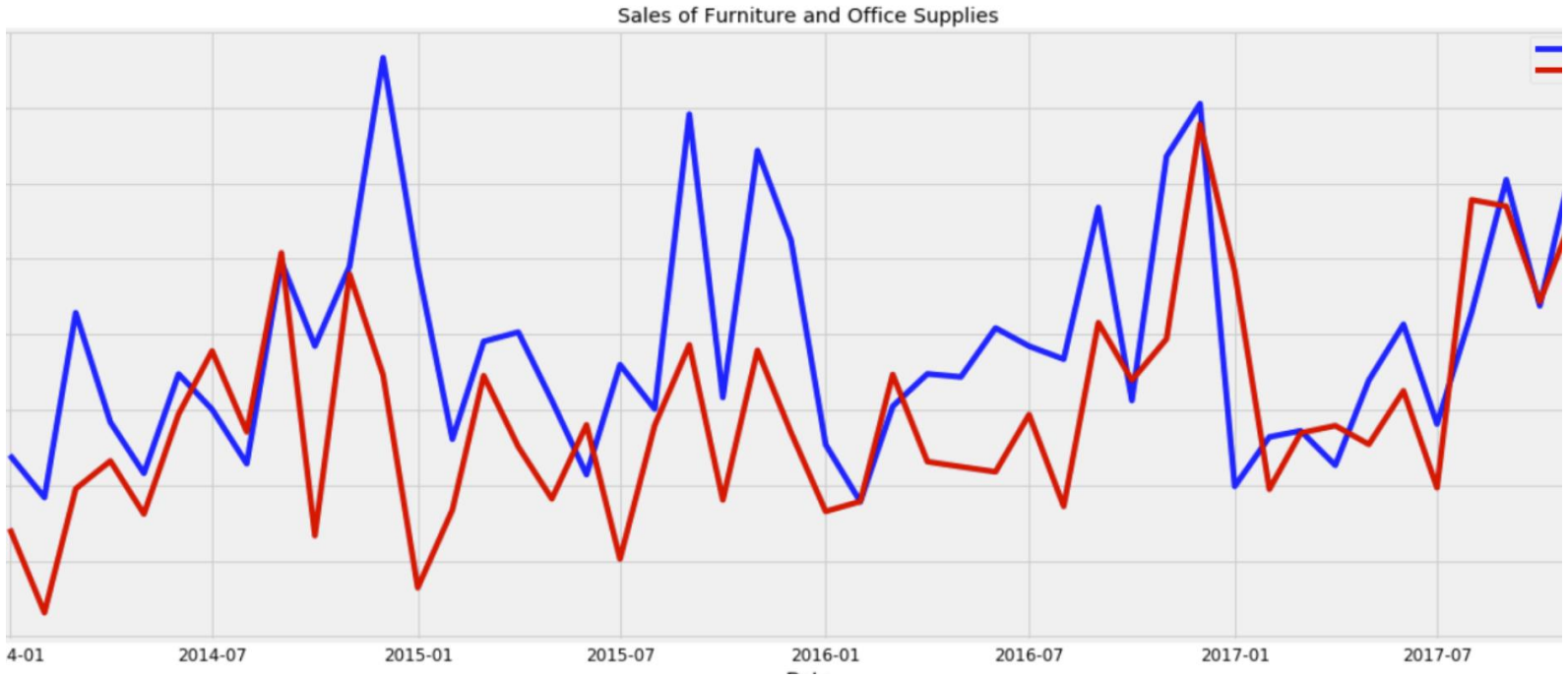
Data warehouses are organized around major subjects, such as customer, product, sales, etc.

Subject-oriented



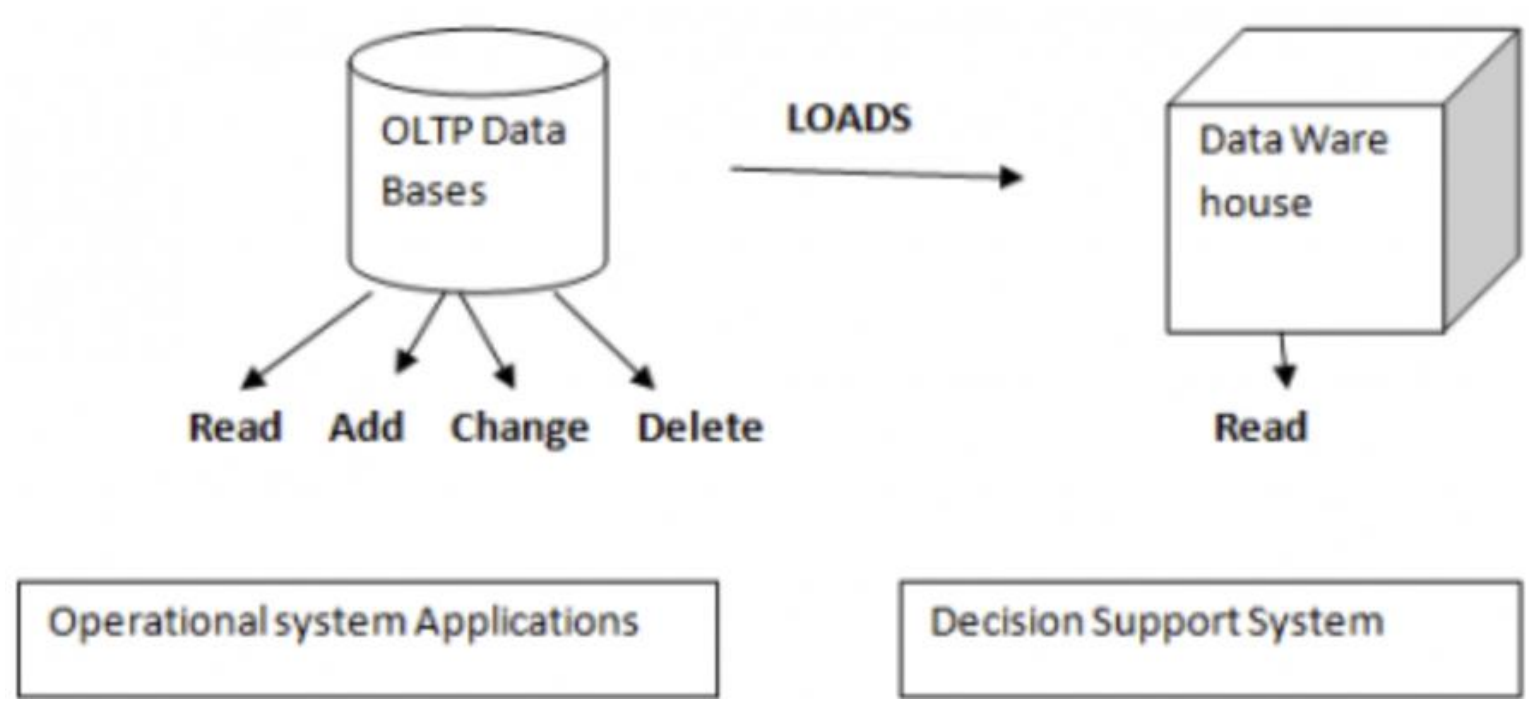
Data warehouses are constructed by integrating data from multiple heterogenous data sources.

Integrated



Data warehouses store historical data.

Time variant

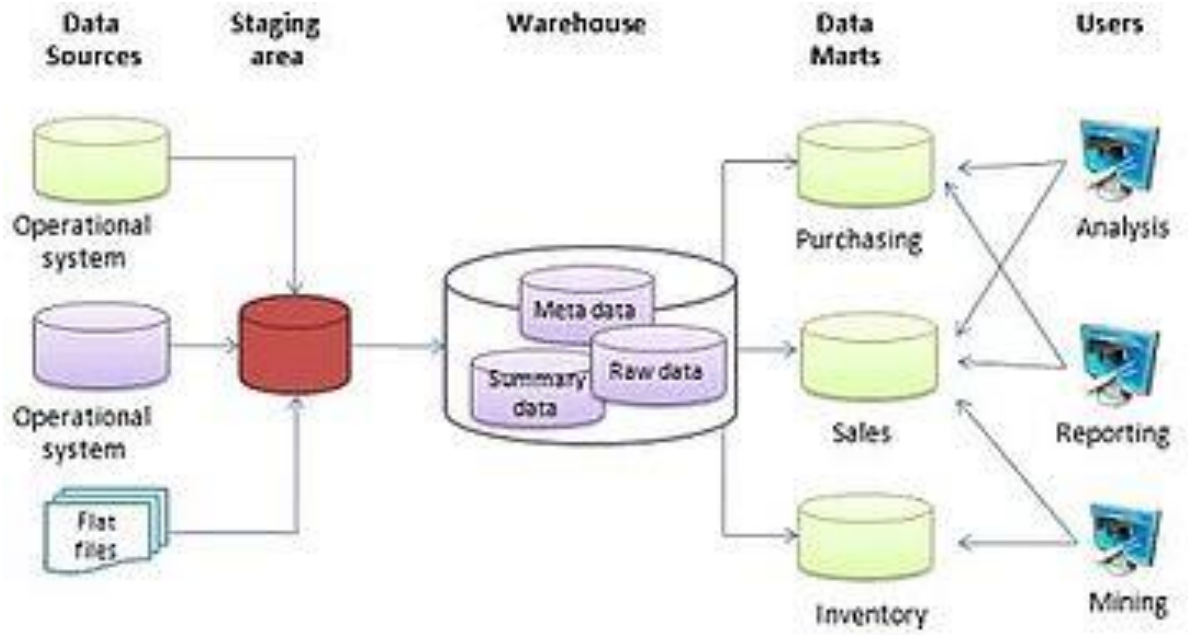


When new data is added to the data warehouse, previous data is not erased.

Non-volatile

Online analytical processing (OLAP) is used for analytics, while online transaction processing (OLTP) is used for transactions.

	OLTP	**OLAP**
users	clerk, IT professional	knowledge worker
function	day to day operations	decision support
DB design	application-oriented	subject-oriented
data	current, up-to-date detailed, flat relational isolated	historical, summarized, multidimensional integrated, consolidated
usage	repetitive	ad-hoc
access	read/write index/hash on prim. key	lots of scans
unit of work	short, simple transaction	complex query
# records accessed	tens	millions
#users	thousands	hundreds
DB size	100MB-GB	100GB-TB
metric	transaction throughput	query throughput, response

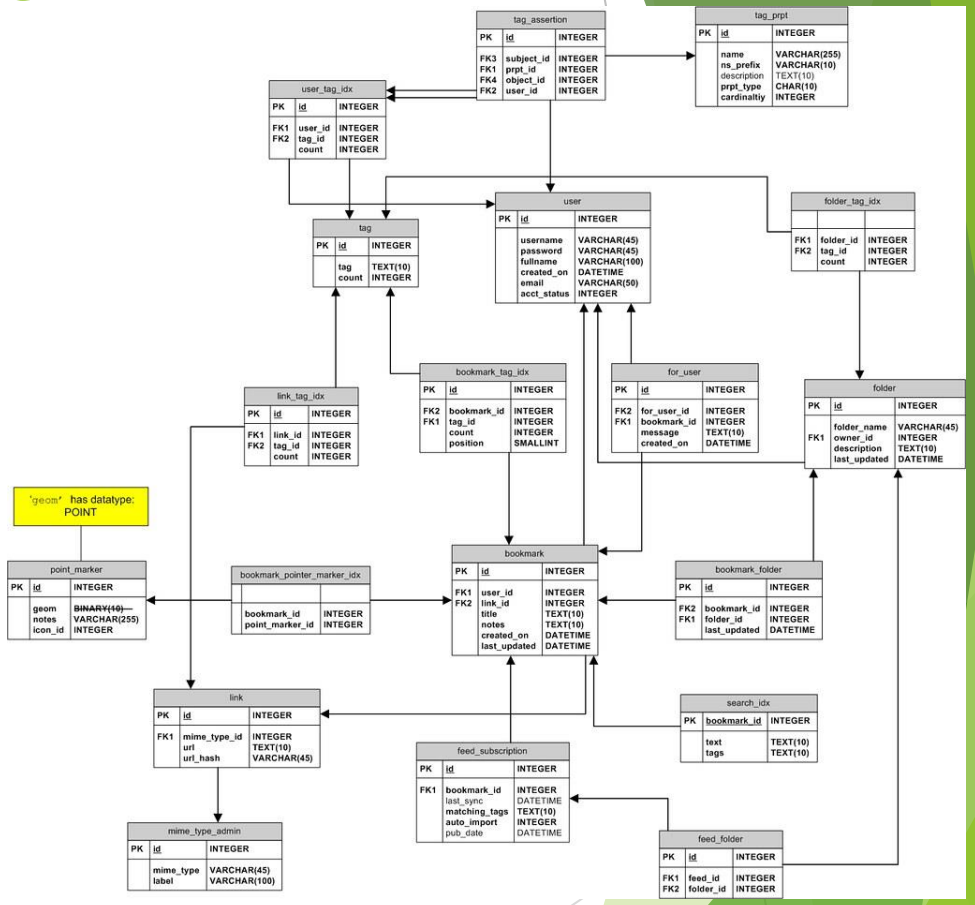


A multi-tiered architecture allows for higher performance and simplicity.

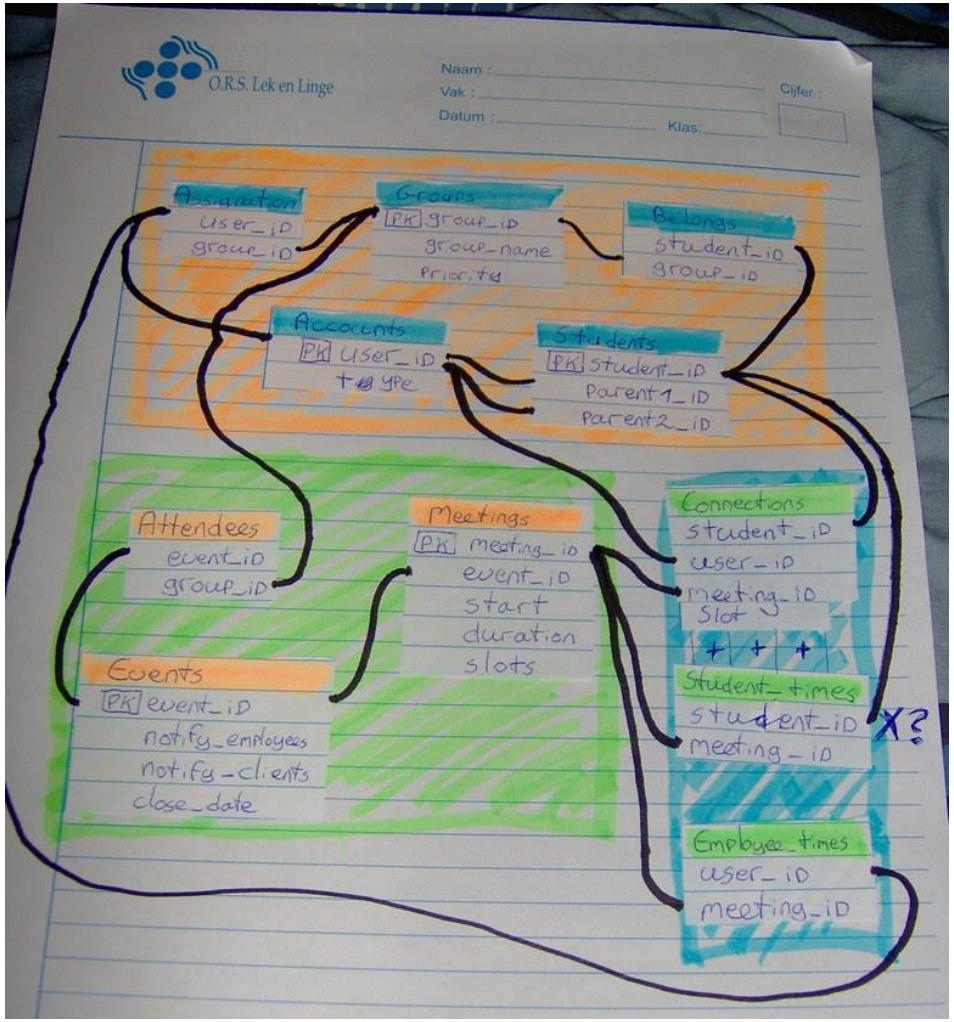
3 Kinds of Data Warehouse Models

- Enterprise Warehouse
- Data Mart
- Virtual Warehouse

Enterprise Warehouses are huge and meticulously-planned



Data Marts form at the departmental level



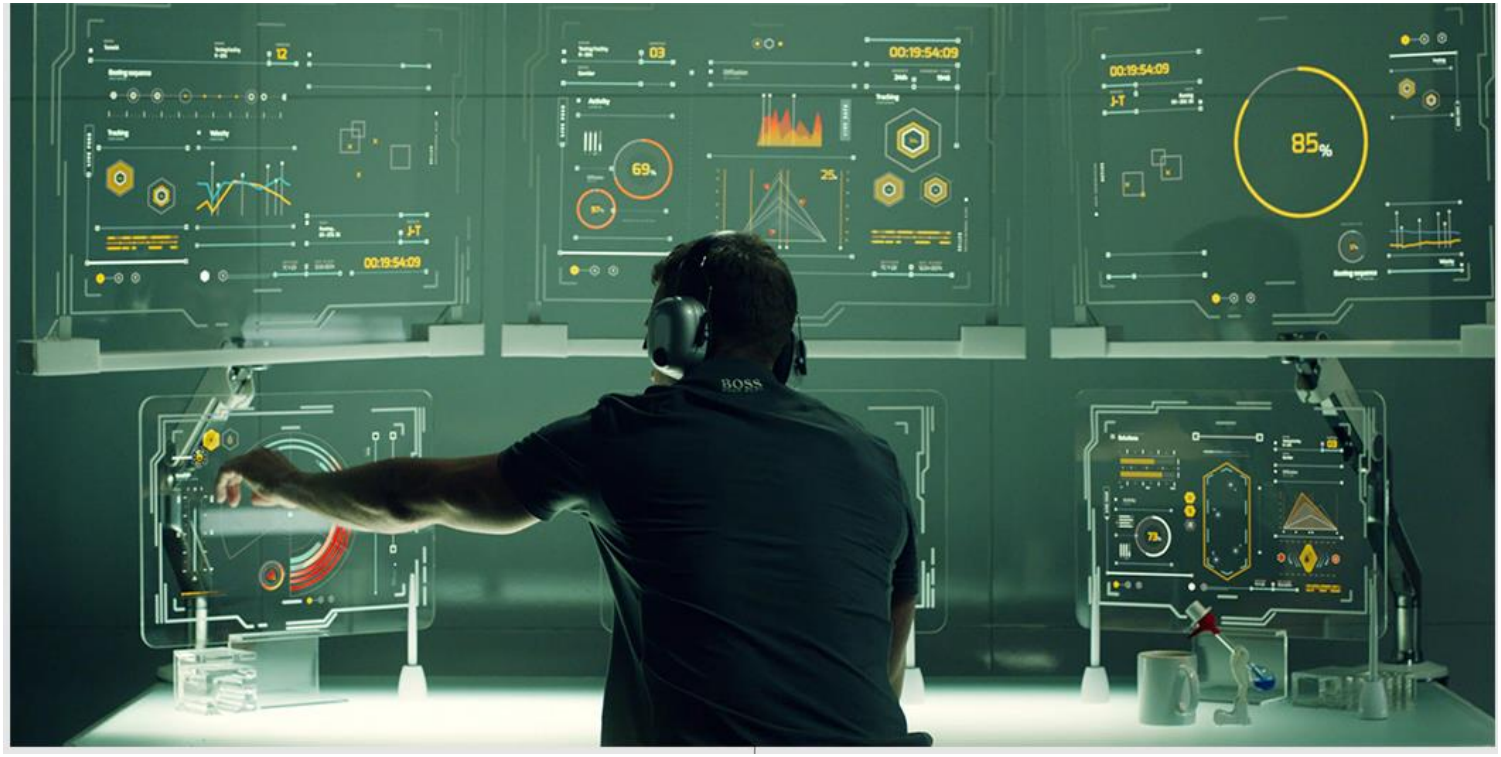
Examples

Eller Full-Time MBA Class of 2021

While our students share a passion and drive for this program and their careers—they come from a diverse blend of backgrounds.

They have nearly five years of professional experience in the workplace before enrolling, and bring an average undergrad GPA of 3.5 to the table. There's a rich mix of academic backgrounds, interests and countries they call home. And they're working together to push each other, support each other and graduate ready to take on the world.

29 average age	**4.3** average years of experience	**15%** military
659 average GMAT	**3.5** average GPA	**40%** female



A Virtual Warehouse is like a data dashboard

Three Data Warehouse Models

https://uair.arizona.edu/content/overview

Scores

Enrollment

Full-Time Equivalent Enrollment

Retention and Graduation

Degrees and Majors Awarded

Tuition and Mandatory Fees

Honors College

College of Medicine

aculty and Staff

Workforce Demographics

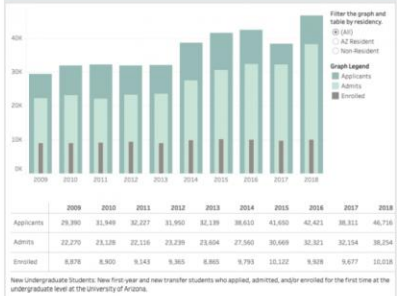
Faculty Demographics

Faculty Tenure

inance

Student

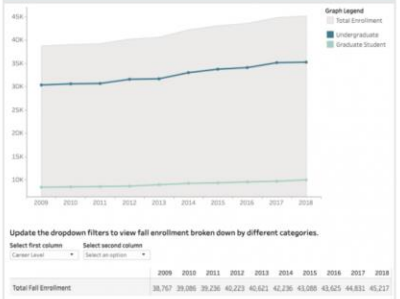
Admissions Pipeline



Fall admissions data is available for new undergraduates, including first-year and transfer students, and for new graduate students. Data can be filtered by residency.

ADMISSIONS PIPELINE WORKBOOK

Enrollment Trends



Student enrollment data is available based on career level and new student status. Data is also available specifically for transfer and international students.

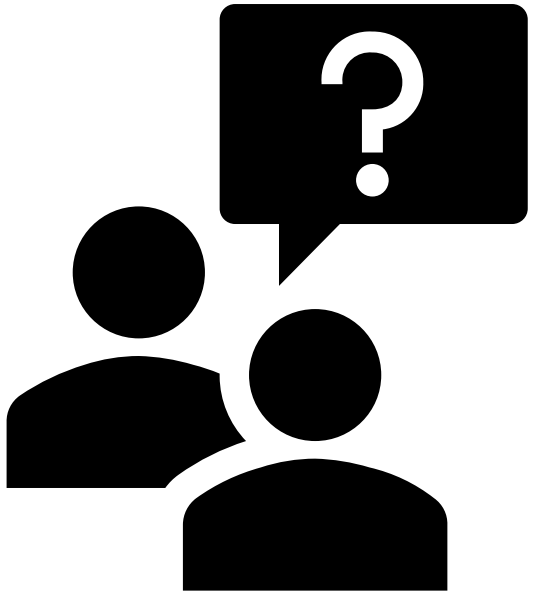
ENROLLMENT TRENDS WORKBOOK

Retention and Graduation

Retention and graduation rates are available starting with the fall 2006 entry cohort. Data can be filtered by gender, IPEDS race/ethnicity, and residency.

RETENTION AND GRADUATION WORKBOOK

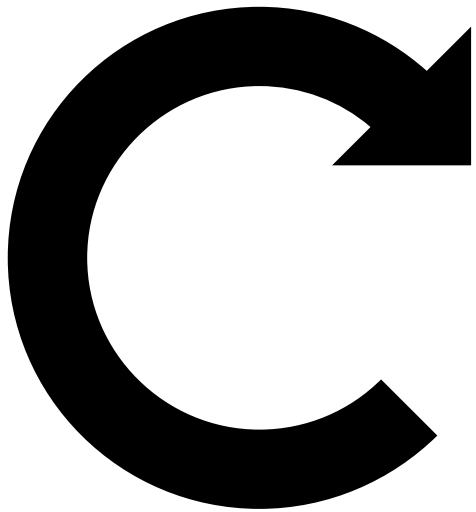
- When do we need it?
- Will this fit with the enterprise?
- How centralized should it be?



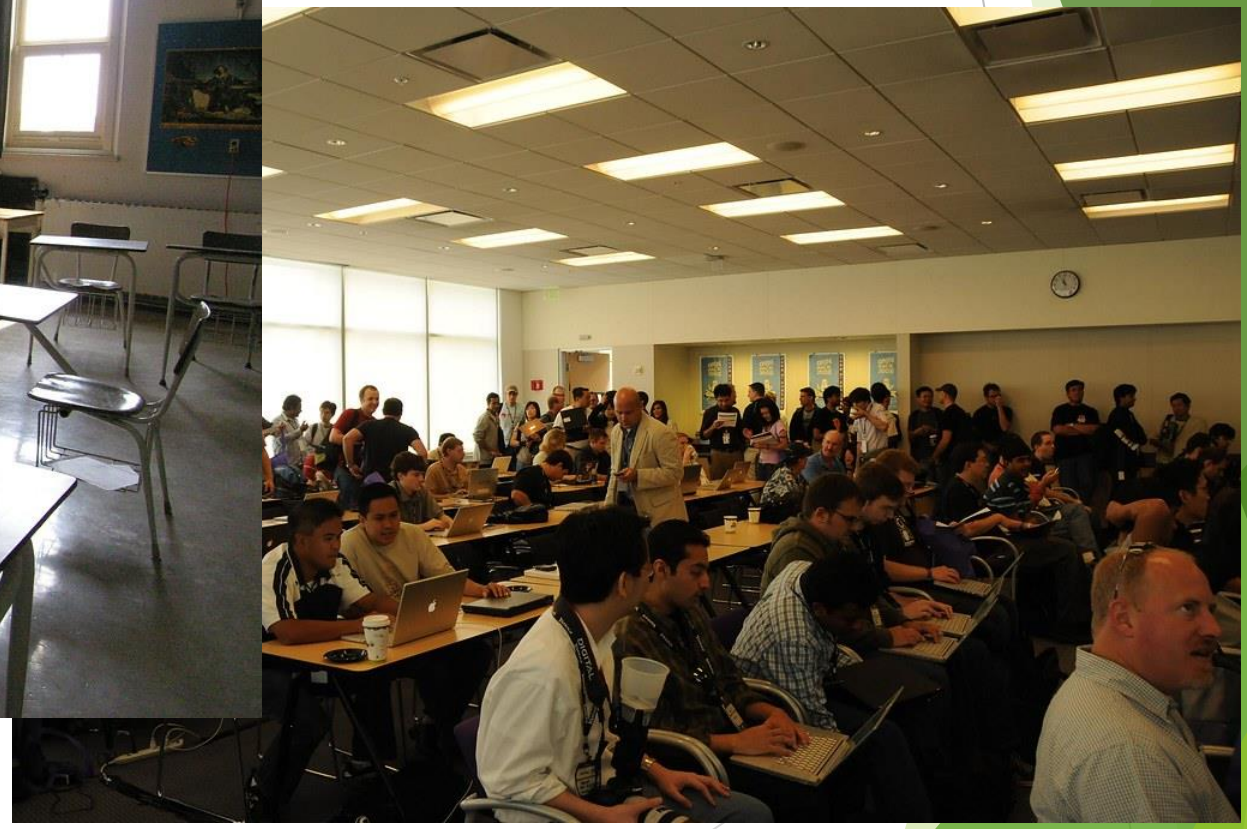
Which data warehouse structure should we use?

Utilities to populate and refresh data:

- Extraction
- Cleaning
- Transformation
- Loading
- Refreshing

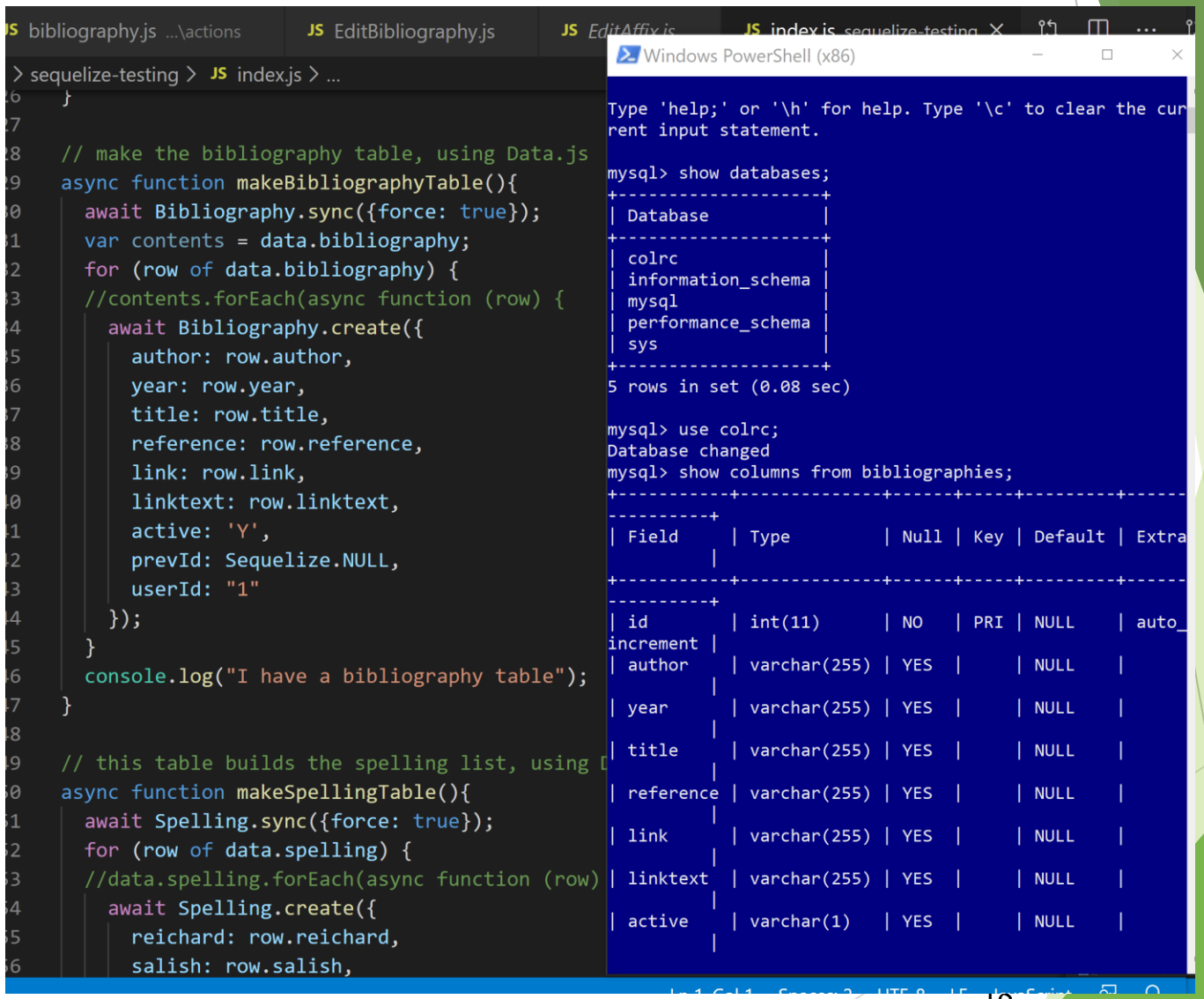


Refresh intervals make a difference.



Extraction, Transformation, and Loading (ETL)

- From JSON to MySQL using Sequelize.



Extraction, Transformation, and Loading (ETL)

```
const bibliography = [
  {
    "id": 1,
    "author": "Barthmaier, Paul T.",
    "year": "1996",
    "title": "A Dictionary of Coeur d'Alene Sa
    "reference": "University of Montana M.A. Th
    "link": "https://scholarworks.umt.edu/etd/8
    "linktext": "here"
  },
  {
    "id": 2,
    "author": "Bischoff, Shannon T.",
    "year": "2011",
    "title": "Lexical affixes, incorporation, a
    "reference": "Studia Linguistica 65.1:1-32.
    "link": ""
  },
  {
    "id": 3,
    "author": "Bischoff, Shannon T.",
    "year": "2011",
    "title": "Formal notes on Coeur d'Alene cla
    "reference": "Newcastle: Cambridge Scholars
    "link": ""
  },
```

Metadata is Data about Data.

3) All imputed income variables have been removed from this version of the dataset. The raw income variables (with ceilings of $200,000 for individual items, and $300,000 for composite items) are still included.

4) B1SMARRS (marital risk) has been recomputed using sum of B1SL7 and reverse coded B1SL8.

5) Created new variable B1SDAYDI (intermediate activities of daily living) that includes one more item than B1SBADL1.

6) Three cases had erroneous scores for B1SPWBU7 (7-item purpose in life).

7) Some chronic pain variables have been correctly renamed:

Old	New
B1SA21E	= B1SA21D
B1SA21G	= B1SA21E
B1SA21I	= B1SA21F
B1SA21D	= B1SA21G
B1SA21F	= B1SA21H
B1SA21H	= B1SA21I.

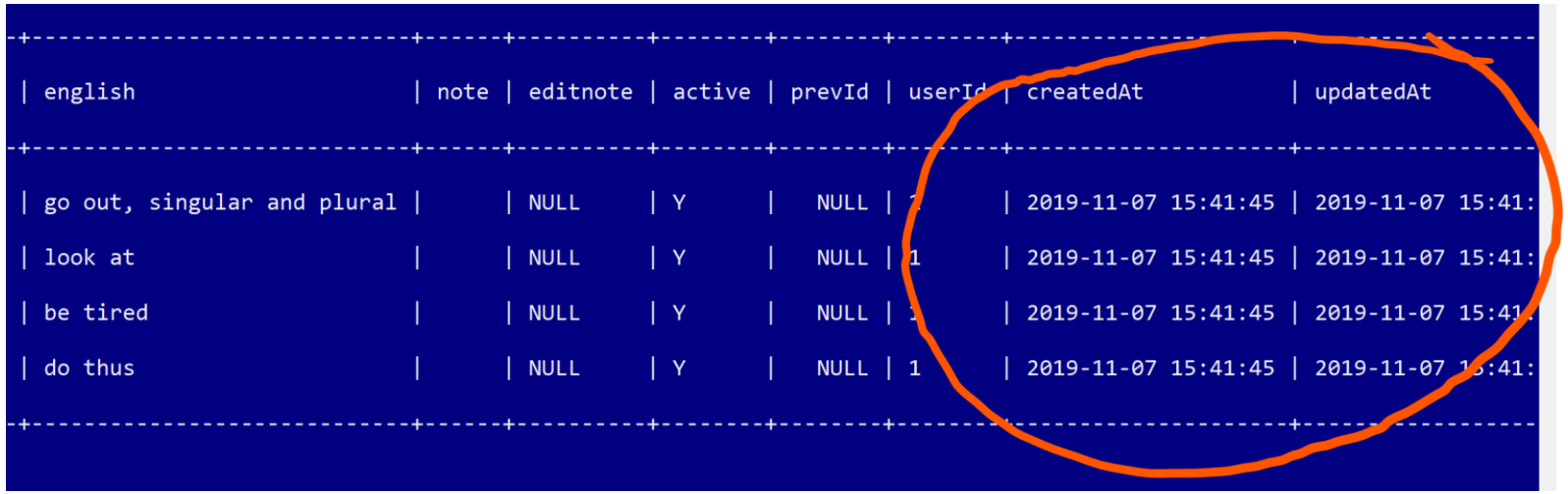
From the readme Included with our dataset

C. What is the Structure of the MIDUS 2 Project 1 Dataset?

This is a rectangular dataset comprised of Phone and SAQ data for 4,963 cases and nearly 2,300 variables. The dataset combines respondents from the Main, City Oversample, Sibling, and Twin samples. The same aggregation of samples exists in the MIDUS 1 dataset. The variable called SAMPLMAJ identifies which of the four subsamples a case belongs to. Variables have been named according to the Short Variable Name (SVN) conventions. All variables include labels to aid interpretation. Value labels have been applied where appropriate. Discrete missing values have also been defined and the following labels applied: DON'T KNOW, REFUSED/MISSING, and INAPPROPRIATE. Details about variable naming and value labels can be found in the naming

Example of a metadata repository: Operational metadata

- Includes data lineage (history of migrated data and thesequence of transformations applied to it), currency of data (active, archived, orpurged)



english	note	editnote	active	prevId	userId	createdAt	updatedAt
go out, singular and plural		NULL	Y	NULL	1	2019-11-07 15:41:45	2019-11-07 15:41:
look at		NULL	Y	NULL	1	2019-11-07 15:41:45	2019-11-07 15:41:
be tired		NULL	Y	NULL	1	2019-11-07 15:41:45	2019-11-07 15:41:
do thus		NULL	Y	NULL	1	2019-11-07 15:41:45	2019-11-07 15:41:

Thank you for your attention!

- Questions?

Chapter 4: Data Warehousing and On-line Analytical Processing

- Data Warehouse: Basic Concepts
- Data Warehouse Modeling: Data Cube and OLAP
- Data Warehouse Design and Usage
- Data Warehouse Implementation
- Data Generalization by Attribute-Oriented Induction
- Summary